Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jakub Růžička
Název práce	Normalizace čísel pro výslovnost syntézou řeči
Rok odevzdání	2023
Studijní program	Informatika
Studijní obor	Informatika se specializací Programování a vývoj software
Autor posudku	Mgr. Ondřej Dušek, Ph.D. Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání	X	X		
Splnění zadání	X			
Rozsah práce *… textová i implementační část, zohlednění náročnosti*	X			

Bakalářská práce Jakuba Růžičky se zaměřuje na preprocessing číselných výrazů pro českou syntézu řeči (TTS). Implementační práce na ní probíhaly ve spolupráci se společností MAMA AI. Ze zápisu číselných výrazů pomocí číslic, jak je bežné v psaném textu, je pro potřeby vyslovení v TTS vygenerovat konkrétní slovní výraz (*153* → *sto padesát tři*, *29. 6.* → *dvacátého devátého šestý* apod.). V češtině je toto ještě komplikováno potřebou použít správnou vyskloňovanou formu slovního výrazu. Navzdory překotnému vývoji v oblasti TTS, neuronovým modelům a běžně dostupným cloudovým TTS s velmi přirozenými hlasy je tento problém stále aktuální – i nejnovější systémy mají s vyslovováním číselných výrazů problémy. Pro specifické aplikace např. v dialogových systémech zaměřených na bankovnictví, dopravní spoje atp. se jedná o zásadní omezení. V češtině byla dosud veškerá řešení buď velmi omezená, nebo proprietární. Kromě toho nebyla k dispozici žádná veřejně dostupná data, na kterých by bylo možné systémy evaluovat.

Autor ve své implementaci zvolil kombinaci pravidlových a statistických metod. Inspiroval se přitom některými publikovanými systémy pro jiné jazyky, ale celé pojetí upravil, aktualizoval a rozšířil pro češtinu. Pravidlová část modelu je založena na vážených konečných stavových transducerech (WFST), kde autor implementoval pravidla pro detekci většiny běžných numerických výrazů (základní i řadové číslovky, desetinná čísla, datumy, časy, míry, peněžní výrazy atd.) a jejich převod na slovní vyjádření, včetně generování více možných variant a zachování morfologické shody. Statistická část má podobu rerankeru, který z více vygenerovaných variant vybírá tu nejpravděpodobnější pomocí předtrénovaných jazykových modelů. Zde autor používá tři různé volně dostupné předtrénované modely pro češtinu s architekturou Transformer v podobě dekodéru, což je momentálně nejmodernější a nejsilnější přístup k jazykovému modelování a přiřazení pravděpodobností. Použití statistických modelů jako rerankeru (namísto generování celého výrazu statisticky) je bezpečné řešení, vhodné pro produkční systémy – vždy se vybírá pouze z variant, které předcházející WFST považuje za možné a korektní.

(pokračování na další straně)

Kromě samotných modelů je důležitým výstupem práce i evaluační dataset, který autor posbíral vlastními silami s pomocí dobrovolníků, ale v profesionální kvalitě (dvojitá kontrola všech anotovaných výrazů). Autor zároveň pro potřeby anotace naimplementoval webovou aplikaci, kterou lze snadno použít pro jiné podobné anotační projekty. Anotovaný dataset je rozdělen na dvě části podle zdroje – zprávy Českého rozhlasu a texty z Wikipedie; obsahuje pouze věty s numerickými výrazy. Celkem se jedná o více než 1800 vět a 3000 numerických výrazů.

Implementované metody normalizace jsou detailně vyhodnoceny na anotovaném datasetu z pohledu přesnosti (tj. přesné shody slovních výrazů s lidskými anotátory). Autor evaluuje několik variant systému (bez jazykového modelu i s různými modely, použitými pro skórování individuálních číselných výrazů i celých vět). Porovnání obsahuje základní baseline z open-source knihovny *num2words* a zároveň i horní baseline (oracle – dosažitelnost konkrétních forem pomocí implementovaných pravidel). S nejsilnějším jazykovým modelem dosahuje systém na poměrně komplexních datech z Wikipedie přesnosti na úrovni jednotlivých výrazů 84% (oproti 78% bez jazykového modelu a pouhým 23% s *num2words*), což je výborný výsledek.

Autor dále evaluaci dělí podle jednotlivých typů numerických výrazů a přidává i chybovou analýzu a ručně provedené srovnání na malém vzorku s jedním z dnešních nejlepších cloudových modelů, Google TTS. Ukazuje, že ve velké části případů označených automatickou evaluací za chyby generuje systém validní varianty, které však nebyly vybrány anotátorem. Většina ostatních problémů jde na vrub omezení pravidel, popř. neočekávaným typům výrazů; obojí je relativně snadno řešitelné. Ve srovnání s Google vychází systém též velmi dobře – sice produkuje více gramatických chyb, ale nemění význam, což je častý problém Googlu.

Zadání a rozsah implementace i evaluace přesahuje běžné nároky na bakalářskou práci; zadání bylo zcela splněno a v některých ohledech překonáno (rozsah podpory typů numerických výrazů, kvalita a rozsah evaluačních dat). Zvažujeme i publikaci výsledků na konferenci, zatím k tomu nedošlo čistě z časových důvodů. Implementaci autor řešil z velké části sám ve vlastní iniciativě a v konzultacích s kolegy z MAMA AI, společně jsme probírali spíš high-level záležitosti. Evaluace a psaní textu pak probíhaly ve velmi úzké spolupráci se mnou, autor pracoval velice precizně a perfektně svůj postup komunikoval. Veškeré mé připomínky jsme vyřešili ještě před odevzdáním práce.

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *… jazyková úroveň, typografická úroveň, citace*	X	X		
Struktura textu *… kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*	X			
Analýza	X			
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

Text splňuje všechny náležitosti potřebné pro bakálářskou práci. Rozdělení do kapitol je velmi přehledné a zahrnuje důležité teoretické pozadí, související práce i podrobný popis implementace, evaluačního datasetu i samotné evaluace a výsledků práce. Související části textu na sebe vhodně odkazují. Práce je psaná velmi dobrou angličtinou, text je dobře srozumitelný.

Příloha práce obsahuje i podrobný uživatelský manuál pro anotaci dat. Dokumentace pro anotační aplikaci, anotovaný dataset i samotný normalizační systém je stručná, ale zahrnuje vše podstatné, takže k ní nemám žádné výhrady. Kód samotný je přehledný a dobře dokumentovaný.

Implementační část práce

		lepší	OK	horší	nevyhovuje
Kvalita návrhu	*… architektura, struktury a algoritmy, použité technologie*	X			
Kvalita zpracování	*… jmenné konvence, formátování, komentáře, testování*	X			
Stabilita implementace		X			

Jak už je zmíněno výše, implementace je velmi dobře zvolena a pro rerankování používá dnešní nejmodernější přístup, tj. transformerové jazykové modely. Kód je logicky strukturovaný, dokumentovaný a velmi dobře čitelný. Čistota kódu je rozhodně nad moje očekávání, autor např. použil Pylint pro konformitu se standardem PEP8 a v celém kódu používá typové hinty pro zvýšení čitelnosti a stability. S instalací, zprovozněním a stabilitou jsem neměl žádné problémy – vše funguje tak, jak má.

Celkové hodnocení Výborně

Práci navrhuji na zvláštní ocenění Ano

Datum: 20. 6. 2023

Podpis